# GEMS/Food Programme

## REPORT OF THE WHO WORKING GROUP ON DATA REPORTING FOR HAZARDS OCCURING IN FOOD (HOF)

**25 June 2010**

World Health Organization

DEPARTMENT OF FOOD SAFETY AND ZOONOSES

***This document is available on the Internet at:***
http://www.who.int/foodsafety/chem/gems/en/index.html

***Copies may be requested from:***

Department of Food Safety and Zoonoses
World Health Organization
20, Avenue Appia, CH-1211 Geneva 27
Switzerland

WHO Library Cataloguing-in-Publication Data

GEMS/Food programme: report of the WHO working group on data reporting for hazards occurring in food (HOF).

1.Food analysis - standards. 2.Food contamination - analysis. 3.Diet. 4.Diet surveys. 5.Data collection - methods. I.World Health Organization. Dept. of Food Safety, Zoonoses and Foodborne Diseases.

ISBN 978 92 4 150068 5 (NLM classification: QU 146)

# Contents

# REPORT OF THE WHO WORKING GROUP ON DATA REPORTING FOR HAZARDS OCCURING IN FOOD (HOF)

## 1. BACKGROUND

The GEMS/Food programme consists of three main areas of activity:

- Data collection for hazard occurrence, food composition and food consumption and sharing information for accurate international risk assessments.
- Training and capacity building at national and regional level for chemical exposure assessment based on Total Diet Studies.
- Development and to dissemination of guidelines and recommendations for exposure assessment methodologies.

WHO is in the process of updating its database format and data reporting procedures to ensure that:

- all information considered essential for estimating dietary exposure is captured;
- the information is accessible in a format that will allow data from different Member States to be compared and combined for the purpose of risk assessment;
- food composition/chemical occurrence data can be combined with available food consumption data for estimating dietary exposure; and
- the database will accommodate both aggregated and individual analytical results.

WHO also wants to anticipate future data needs for conducting risk assessment including:

- the ability to better quantify acute risks of exposure;
- the use of probabilistic approach for chronic exposure assessment which requires the full distribution of occurrence data instead of just mean/median and upper percentile;
- the ability to conduct a risk/benefit analysis, which requires information on nutrient intakes; and
- the possibility of including in the database information on microbiological hazards in foods.

WHO asked the Working Group (WG) members to provide advice on:

- the type of information that is required to accomplish the goals above; and
- ways to reduce the work load of those responsible for reporting/entering the data, including the staff of both Member countries and WHO.

Using the current GEMS/Food database structure as a starting point, the WG members identified the data fields that each considered to be essential information, and provided comments on field formats and content. They also provided comments on current data submission procedures and suggested ways to improve and enhance the process.

## 2. TERMS OF REFERENCE FOR THE WORKING GROUP

The basis for estimating the dietary exposure to food hazards is to combine data on food consumption with levels of occurrence of the hazards in the corresponding food categories. The GEMS/Food programme aims to support the dietary exposure assessment procedure by way of a network of national/regional institutions that collect data and then report the data to the WHO GEMS Food database. The GEMS/Food database includes data on food consumed all around the world and aggregated in 13 diets. The current database also compiles levels of occurrence for chemical contaminants.

a. Data needs for food composition and chemical/microbiological occurrence
Crucial issues in assessing food safety for both the risk assessor and the risk manager are, on the one hand to have access to available data on food composition and chemical occurrence (and food consumption?) and on the other hand to share their own information with partners in other Member States. This information should be available and accessible in a format allowing comparison and combination of data from different Member States for the risk assessment purpose.

The overall objective of collecting data on hazard occurrence and food consumption in a shared format is to support the exposure assessment process. Because data collection requires long term programming to prepare hosting specific formats, it is necessary to anticipate the future changes and evolutions of the whole risk assessment process.

The main evolutions which can be expected in the next years are: (1) **A better quantification of the acute risk** both for chemical and microbiological hazards. (2) **A probabilistic approach for chronic exposure assessment**. The full distribution of occurrence data should be needed instead of the central tendency and high percentiles of the distribution. Considering the whole distribution will imply the use of statistical modelling in certain occasions in particular to deal with non-quantifiable analytical results (results below the LODs). (3) **A risk/benefit evaluation** including a balance between the exposure to hazard and possible benefits of nutrients intake. Therefore a database on nutrient composition of food should be developed or shared with other WHO departments as recommended by the FAO/WHO technical workshop on nutrient risk assessment (Rome, 2-6 May 2005).

b. Terms of reference for the working group
The working group should advise WHO on the most appropriate format for data reporting in the GEMS/Food database taking into consideration

- The current and future needs for exposure assessment, in particular those listed in paragraph 3 above.
- The existing food classification systems and in particular the Codex Alimentarius classifications (Appendices 1 & 2).
- The optimal level of food classification/aggregation to allow combining data on chemical occurrence and nutrient composition with food consumption data? (and/or data reported by different Member States?).
- The feasibility in term of workload for data providers.
- The possibility for reporting both chemical and microbiological hazards.

The working group should recommend the information that should be reported and propose draft-templates for data submission by Member States. These templates should be tested in some Member States with the help of the WHO Collaborating Centres. The web-system for reporting as well as the management tools for the databases should follow the rules of the current WHO IT system and in particular should allow exchanges between WHO departments.

## 3. PROJECT SCOPE

The consensus of the WG was that the project scope was far-reaching and perhaps ambitious given the short timeframe for providing comments and preparing this report. It was suggested that the long-term goals of redesigning a data system might best be achieved by taking a stepwise approach. There was general agreement that the project should focus initially on making some modifications to the current data format, by addressing issues regarding food description and classifications, and by streamlining the data submission process. Expanding the database to include other types of information (e.g. microbiological and nutrient data) and database capabilities (e.g. calculating exposures) could be phased in over time. Nevertheless, recognizing the importance of the overall goal of designing a more inclusive data system, the WG has provided recommendations on both the immediate and longer-range goals and components of such a system.

One WG member suggested that a logical starting point for reviewing the database content was to understand and consider how the GEMS/Food data have been used in the past and will likely be used in the short-term. This would provide some guidance as to the level of complexity required of the database and to determine whether specific information would be critical or non-critical.

If the GEMS/Food database has been used primarily as a source of occurrence data to assist with dietary exposure assessments conducted by JECFA and by JMPR, the exposure assessments that are undertaken for this purpose are often simple and do not require complex data sets. Estimating dietary exposure requires linking the GEMS/Food occurrence data to food consumption data. When the JECFA or JMPR assessments are conducted using occurrence data from different countries or regions, often the only consumption data available have been the WHO Cluster Diets (previously the Regional Diets), which would not allow sophisticated modeling techniques to be undertaken. In that case, the data capture requirements for hazard concentration data should be commensurate with the consumption data to be used in the assessment.

Note that consideration of food consumption data in relation to the GEMS/Food system was not part of the Terms of Reference for this WG, as this topic is being reviewed by a separate Working Group. Nevertheless, several comments related to food consumption data have been included in this report particularly in relation to food classification systems since this is the means for linking occurrence data and food consumption data for estimating dietary exposure.

More specific comments and suggestions follow.

## 4. HARDWARE, SOFTWARE & FILE FORMAT CONSIDERATIONS

File/data transfer format

There were several comments supporting the use of a standard file format such as XML or Excel. FSANZ is considering using the XML standard as its data transfer format. It was mentioned (though not confirmed) that the European food composition network, EuroFir, is also planning to use XML. Adopting XML would allow WHO to create an XML Schema specifying what data is required in GEMS/Food. This is conceptually similar to creating a template spreadsheet for data upload, but is much more robust. The data transfer isn't affected by spreadsheet formatting and data providers (as well as WHO) can use software to automatically check that the data submission conforms to the schema's validation rules (i.e. any within field or referential integrity [between field) constraints such as field length or data type). Briefly, some of the advantages of adopting XML are:

- It is becoming a universal data exchange format in other fields.
- It is well supported by nearly all mainstream operating systems, databases and related software (e.g. laboratory management systems).
- It is more robust and flexible than delimited files or spreadsheets.
- XML Schemas allow organizations to clearly specify how they wish to receive data.
- An XML file can be validated against a schema automatically.
- Once a process to output or input XML data has been setup or coded, it can be used repeatedly, facilitating more frequent automatic data exchanges.
- The technology reduces manual work at both ends, saving a large amount of time and reducing the chance that errors will be introduced to a dataset.
- XML files are text based and are human-readable if necessary.

Excel is also a good option for this application and offers some advantages over XML. Excel is widely available, and a spreadsheet template can be packaged with functionalities whereas XML is a plain text file that requires software to provide the functionalities. Excel would enable data submission by countries that do not have the technical expertise or support necessary for using XML.

Desk-top versus web-based application

One major drawback of the current OPAL system is that it must be locally installed on the data provider's computer. This is often problematic for several reasons:

- Administrative rights are often required to install software.
- There may be installation problem related to the operating system.
- The update procedure is complicated.

A better solution would be to create a centrally updatable, web-based application with an OPAL type front-end. An application (OPAL or similar) could still be used in areas where internet access is limited or if the user prefers to install the system locally.

If a web application is developed, it must be able to work on different web browsers (e.g. Firefox, Internet Explorer, Safari).

A web application would also require that consideration be given to appropriate security levels. Different levels of security can be created for specific users (e.g. reader, uploader, administrator, etc.) who will have access only to the appropriate features. Also, the system should not have features commonly blocked by security systems, such as pop-ups.

Other system capabilities

Other system features that would be very useful to the end user include:

- access to standardized calculation instruments; and
- report generation capability to obtain standardized output to be used for communication of results.

## 5. HARDWARE DATABASE CONTENT – CURRENT AND FUTURE

The Working Group was asked to respond to the following three questions related to database content, considering both current and future need:

1. In the current GEMS/Food database, which fields at do you consider critical (essential) versus non-critical (good to have but not essential)?
2. What are the recommended properties of each data field (e.g. text versus numeric, length of field, controlled language/text versus free form text)?
3. What other types of information are needed for conducting risk assessments, especially anticipating the future needs identified by WHO (refer to BACKGROUND section above).

The current GEMS/Food system comprises two types of occurrence data: aggregated data and individual data. The fields for each type of data are listed in the table below. Appendix A includes more information about each of the data fields, including field size, applicable codes and other explanatory notes.

Critical versus non-critical information

There was general agreement among the WG regarding fields considered to be critical (i.e. that the information is essential for ensuring high quality data and should be required for data submissions). One member considered all of the fields to be critical. From another perspective, one member noted that it is important to recognize that, for many countries, incomplete data sets and resource restrictions are a fact of life, therefore the GEMS system should be as simple as possible to use with as few as possible mandatory fields. Critical fields should be identified in the user manual (e.g. in the Appendix of GEMS/Food Database Fields).

The following table summarizes the majority opinion of the WG with regard to the critical nature of each database field.

| GEMS/Food Database Field | In database for aggregated data | In database for individual data | Critical vs not critical |
|---|---|---|---|
| Serial no. of the record | ● | ● | Not critical |
| Creation date of record | ● | ● | Not critical |
| Country code | ● | ● | Critical |
| Food identifier | ● | ● | Critical |
| Food origin | ● | ● | Not critical |
| Sampling period (or date) | ● | ● | Critical |
| Sample representativeness (or reliability) | ● | | Not critical |
| Laboratory identification | | ● | Not critical |
| Number of contributing laboratories | ● | | Not critical |
| Analytical quality assurance | ● | ● | Not critical |
| Contaminant | ● | ● | Critical |
| Dimensions of results | ● | ● | Critical |
| LOD | | ● | Critical |
| LOD min | ● | | Critical |
| LOD max | ● | | Critical |
| LOQ | | ● | Critical |
| LOQ min | ● | | Critical |
| LOQ max | ● | | Critical |
| Results based on (fat/dry) | ● | ● | Critical |
| Result | | ● | Critical |
| Number of samples | ● | | Critical |
| Number of samples less than the LOQ or LOD | ● | | Critical |
| Range - min | ● | | Not critical |
| Range - max | ● | | Not critical |
| Mean or best estimate | ● | | Critical |
| Mean - lower bound | ● | | Not critical |
| Mean - upper bound | ● | | Not critical |
| Median or best estimate | ● | | Not critical |
| 90$^{th}$ Percentile | ● | | Not critical |
| Standard deviation | ● | | Not critical |
| Status of data (confidentiality) | ● | ● | Not critical |
| Remarks/references | ● | ● | Not critical |

More specific comments about the data fields are provided below. Refer to Appendix A for specifics regarding field size and format, applicable codes, and explanatory notes for the current GEMS/Food database.

General recommendations:

Field length

- A suggestion was made that it is no longer necessary to be overly restrictive on field length; field length restrictions should only be used if essential.

Date format

- Different date formats are used in different fields, e.g. creation date is 04/04/2001, and sampling date is 04-Apr-2001. One format should be used throughout.

- GEMS/Food should be able to convert various date formats to a set format.
- Using a 4-digit year rather than 2-digit year is recommended in all cases.

Use of codes rather than text

- There are advantages and disadvantages associated with using either codes or text, so it is recommended that thoughtful consideration be given to this issue.
- Reliance on codes to identify countries or laboratories or contaminants as in the current system saves storage space but this should not be necessary as field length restrictions are generally no longer an issue from an IT perspective. Storage space is now cheaper and search algorithms, such as full-text search, have become more useful with increases in processing power.
- Using codes ensures more consistency in data entry and allows for information to be reported without having to be translated to a common language (e.g. English). Consistency of data entry is critical for working with large data sets such as GEMS/Food and for ensuring that data can be correctly and efficiently compiled. Using codes does, however, require the use of 'code breakers' and additional time is needed to translate the codes.
- As for allowing free-text entries, this eliminates the need to translate text to codes or vice versa but it allows for different spellings of the same word or phrase, which can make working with and compiling data difficult and more time-consuming than code translation.

Blank entries

- To the extent possible, blank cells should be avoided. When a cell is blank, it is not possible to know if the data provider forgot to give the information or if it means no information is available. In the case of analytical result, a blank cell can be indicate a missing value or 'no detected' value. Controlled terminology should be used to indicate if the information is 'unknown', 'not applicable', 'not quantified' or 'not detected'.

Storage of images/photographs

- A facility to upload photographs of sampled foods could be useful, or to upload PDFs of detailed reports. Visual information can be useful if the verbal description of a food is insufficient or unclear, and the uploaded files can be stored for traceability.
- Hyperlinks are also a useful feature but suffer the risk of the document being moved and therefore not available in the future.

<u>Comments on specific fields</u>

Serial no. of the record

- Additional clarification on this field would be helpful (i.e. is this number generated by the country submitting the data or by WHO?).
- Auto fill options for country names would be useful.

Creation date of record

- The same format should be used in all date fields.

Country code

- Auto fill options for country names would be useful.

Food identifier

- The current ‘food list’ screen is too simple; the code hierarchy can't be seen.
- Represent the food classification as a hierarchy (alphabetical order).
- Create a system that can help to search within the food classification.
- Provide the food classification in a file format that can be exported/downloaded by users.
- The system should allow multiple coding systems, particularly if the system is expanded for hazards beyond contaminants.
- Field length restriction should be lifted as some other coding schemes are more than 6 digits.
- For foods that aren’t included in the existing classification systems, there could be a facility to develop an interim identifier, which could then be checked by the GEMS Food administrator. Additional descriptive information about the food could be included in a separate field.
- The ability to store free text information may be useful in terms of providing detail about the food that has been analysed (such as scientific name or names in different languages); this could be as an optional field.

* Note that more specific recommendations regarding food classifications are provided in the section **FOOD CLASSIFICATION SYSTEMS** below.

Food origin

- This field should be more clearly defined and should allow entry of more than one country, as processed foods often contain ingredients from more than one country. For example, it is common for products sold in Australia to indicate they are made from ‘Australian and imported ingredients’ but the importing country is rarely identified in mixed foods and the particular ingredients that are imported are not identified.
- Another option would be to simplify the responses to ‘imported’, ‘domestic’, ‘mixed origin’ or ‘unknown’.
- This information may not always be available so this field should not be mandatory.

Sampling period (or date)

- Same date format should be used in all date fields; note that two different formats are used for aggregated and individual data.
- If a format with only the month and year are used, GEMS/Food should be able to convert various date formats to ensure that the leading zero for the month is included.
- For aggregated data, consider creating separate fields for beginning and ending periods to allow for easier and consistent sorting of data.

Sample representativeness/reliability

- Some guidance on the meaning of ’statistical’ in this context would be helpful.
- It is unlikely that sample results reported to GEMS/Food would be based on a statistical sampling method but the data could still be the result of random rather than

targeted sampling.  The latter information is considered highly desirable since mixing results of random and targeted samples can bias the final estimate.

Laboratory identification

- Importance of information about analytical method and laboratory would depend on the nature of the risk assessment - and if one wished to contact the lab doing the analysis, and whether one needs to compare analytical methodologies.

Number of contributing laboratories

- How is this information to be used?  Could it be omitted?
- This information does not affect the estimate of exposure but the uncertainty around the estimate and, in some cases, the interpretation of the assessment by risk managers.

Analytical quality assurance

- This information is desirable but not essential.
- This information does not affect the estimate of exposure but the uncertainty around the estimate and, in some cases, the interpretation of the assessment by risk managers.
- Add an option for 'Don't know'.
- It might be useful to include a new field that identifies specific proficiency programmes in which labs have participated.

Contaminant

- Auto fill options would be useful.
- Should include chemical form where applicable.
- Contamination identifier list is limited and not updated regularly to reflect emerging contaminant (e.g. melamine, Auramine O). Recommend allowing free-form text for entry of new contaminant. Represent the food classification as a hierarchy (alphabetical order).
- Create a system that can help to search within an international catalogue of unique contaminants codes.
- Provide the list of contaminants in a file format that can be exported/downloaded by users.

Dimensions of results

- If the database is to be expanded to include other types of data (microbiological, nutrient, et al.), appropriate units need to be included.
- Ability to record data in scientific notation would be required for microbiological data.

LOD/LOQ
LOD min/max
LOQ min/max

- There was general agreement that LOD and LOQ are critical fields.  This information is crucial both for samples below the respective limit as well as for those with quantified values as this information decreases uncertainty in modeling and prevents bias that could be introduced if data with high limits are excluded.

- This is also critical information when assessing aggregated results, which should always include an explanation of how non-detects were treated (i.e. what value was assigned to them).

Results based on (fat/dry/as consumed/as is)

- A new text field should be added to capture a detailed description of the food as analysed, e.g. whether it was based on edible portion only or if it included inedible portions such as skin, bones, peel, etc.
- Similarly, it would be useful to capture information on whether it was cooked or raw and how it was cooked.

Result (for individual data only)

- As noted above under 'General recommendations', blank cells should be avoided. When a cell is blank, it is not possible to know if the data provider forgot to give the information or if there was no value to report. Controlled terminology should be used to indicate if the information 'not quantified' or 'not detected' so that it can then be related to the LOD or LOQ.

Number of samples

- Essential information.

Number of samples less than the LOQ

- The majority of WG members felt that this is essential information, though others suggested that it not be mandatory since it may be unknown by many countries and should not deter them from submitting data.
- Add 'or LOD' to the descriptive field name, or change the field descriptor to "Number of samples with quantified results".

Range - min
Range - max

- Aside from suggesting that these are non-critical fields, no specific comments were provided.

Mean or best estimate
Mean - lower bound
Mean - upper bound
Median or best estimate

- Some who have submitted data mentioned that the guidelines referenced in the 'Explanatory notes' (Appendix 5) are unclear and daunting and may deter some countries from submitting data. A simpler set of instructions would be helpful.
- As noted above under 'General recommendations', blank cells should be avoided. When a cell is blank, it is not possible to know if the data provider forgot to give the information or if there was no value to report. Controlled terminology should be used to indicate if the information is 'unknown', 'not quantified' or 'not detected'.
- There should be a separate field to report specifically how censored data were handled in calculation of means and medians. Perhaps the techniques described in Appendix 5, as cited in the Explanatory Notes for this field (refer to Appendix A of this report) could be identified as named options in a drop down list.

90th Percentile

- Should additional fields be added for reporting other percentiles (e.g. 2.5th, 97.5th)? For nutrients, the lower percentiles are often of most interest in terms of assessing adequacy.

Standard deviation

- Avoid having blank cells by including a possible response of ‘not applicable’ or ‘not available’.

Status of data (confidentiality)

- Comments on this field were mixed. Some felt this information was critical, others felt it was not helpful.
- The point was made that once the data are submitted they are no longer confidential. Perhaps a more complex confidentiality policy is needed and the system should include some way to limit the use of or access to designated data.

Remarks/references

- No comments were provided on this field per se, but there were several suggestions for adding fields to capture additional information. These are provided below.

<u>What other types of information are needed for conducting risk assessments, especially anticipating the future needs identified by WHO?</u>

With regard to the current GEMS/Food database, suggestions were given on additional information that would be helpful:

- Due to the issues related to the food classifications, we may want to ask data provider to enter the original food name and contaminant name. This would allow tracking of mapping and coding errors.
- A field to show whether the samples were composite/aggregate samples or not, and if they were, the number of samples making up the composite sample.
- As noted above, a field should be added to describe whether the sample included the edible portion only or if it included inedible portions such as skin, bones, peel, etc.
- A field should be added to report the analytical method used. This could contain a method reference (such as AOAC 27.21) or a simple descriptive term (such as ‘colorimetric’). This would be useful for those chemicals whose identity is defined essentially by the method of analysis (which can be the situation for some nutrients).
- An additional optional field of ‘Analysis date’ could be considered. This may be useful when there has been a substantial time gap between sampling and analysis that may affect the level of the analyte, for example the level of a labile nutrient. It may be relevant in some cases to capture storage conditions of samples between sampling and analysis, again as an optional field. For example it might be relevant if one were looking at packaging migrants to know that samples had not been stored at the laboratory in a container that could contaminate the sample.
- A field to report recovery information on the analysis, i.e. what is the recovery, is it corrected, etc. This should be an optional field.

Following are general comments and suggestions of other fields to be included if the database were expanded to include microbiological data and nutrient composition data. (See also Attachment B for a list of nutrients recommended for consideration by one member):

- In order to capture hazards beyond contaminants and pesticide residues, considerable redesign/extension of fields will be required. One option might be to progress the project in a stepwise process, for example by extending the system to one group of hazards at a time (for example, extend it to cover packaging migrants, then microbiological hazards, then nutrients).
- Incorporating data on multiple hazards into a single system could result in a very complicated system. One possibility would be to have two or three different data reporting systems that could communicate via a single food coding scheme and food consumption data system. Alternatively, if one single system is to be used, it should have different windows for each type of data.
- It would be very helpful to capture information on packaging materials for future risk assessments. The Langual language system developed in Europe (see http://www.langual.org/) includes a lot of useful field information – not all of which may be needed from the risk assessment perspective but they could form optional fields.
- Other groups of data that could be captured, depending on the scope of this project, are the 'recipes' that are used to apportion mixed foods into their component ingredients and the translation/mapping/coding systems used to group foods for different purposes.

Several suggestions related to **microbiological data and risk assessment** were provided:

- There would need to be new fields added, for example to cover time, temperature, storage atmosphere, strain of bacteria etc. The WHO guidelines on microbiological risk assessment probably identify the parameters that must be captured
- For information on microbiological hazards in foods, the stage at which the data collected and type of unit could be added. As stated in MRA series No. 7: Exposure assessment of microbiological hazards in foods, estimates of likelihood (prevalence) and level of contamination (number) should be correspond to the stage where assessment commences (e.g. farm) and the type of unit (e.g. live animal, carcass, subproduct or product) will be determined by this starting point. Secondly, in some microbiological risk assessments, description on the identity of pathogens such as subtypes and genotypes may be required. On the other hand, information for the evaluation of foodborne antimicrobial resistance may be required in the future.
- Other suggested fields include:
    - state of sample: refrigerated/room temp/frozen;
    - typing mechanism: PCR based/pulse field based; and
    - qualitative results: pos/neg (in addition to units for quantitative results).

For assessments involving **nutrients:**

- A comprehensive list of nutrients to consider for inclusion in GEMS/Food was provided by one WG member and is included in Appendix B.
- Fields would be required on whether or not the foods were fortified, cooking method, variety etc. The Langual system covers these aspects in great detail.
- For nutrient intake, other information related to food preparation (e.g. cooking methods, yield/conversion factors) would be helpful for nutrient risk assessment.

- Furthermore, low percentiles of consumption and nutrient intake would also be provided when retrieving the nutrient intake data from the system.
- Including nutrient composition and food consumption data will be a very difficult task, mainly because of the food classification system is quite different from the current GEMS/Food classifications.

For **conducting acute risk assessments**, it will be important to define which chemicals or food components could be of health concern for acute exposure, to estimate acute reference doses, and to develop methodologies for acute exposure to these compounds. In any case, when raw data are not available, a better quantification of the chemical acute risk will probably require consumption and incidence data at higher percentiles (e.g. 95th, 97.5th, 99th, 99.9th). We only have consumption data at international level at higher percentiles for a few commodities, which are used for pesticides.

Information to support the principles of international estimates of short-term intake (IESTI) should be included in the database.

For **risk/benefit evaluation**, it might be best to concentrate on a few issues to start with (mercury/fish) when expanding the system.

If the GEMS/Food system is to also support dietary exposure assessments by incorporating **national consumption data** within the same data system, then a major redesign of the system will be required. Confidentiality of national consumption data must be considered since many countries do not allow the data to be used outside this country without specific permission.

## 6. FOOD CLASSIFICATION SYSTEMS

Two Codex food classification systems (for raw commodities and for processed foods) are used in the current GEMS/Food database. The WG was asked how these might be linked so that all data can be utilized to estimate dietary exposure (e.g. inclusion of processing factors and yield factors).

The WG provided a number of comments related to this question but also to the need to expand the food classification systems to be more descriptive and inclusive. Some comments also referred to food consumption data, since the food classification system is the means for linking occurrence data to food consumption data for estimating dietary exposure/intakes.

<u>How might the two food classification systems for raw commodities and processed foods be linked so that data from both can be used to estimate dietary exposure?</u>

- One WG member was unsure whether the systems need to be linked as they are trying to achieve different things. In their national system, all foods have a code derived from the national food consumption surveys. These foods may then be assigned a raw commodity/processed food code (quite possibly both) if they fit directly into one of the existing classifications, or they may be assigned a recipe, with the recipe ingredients assigned to one of the coding systems; this is a complex and

time consuming step. They have tried to simplify the process but have not identified a way to avoid the dual coding and recipe system.

- Inclusion of processing factors and yield factors could be useful assuming this information is a component of the GEMS/Food system rather than relying on g data providers to gather and include this information with their data submissions. It would be helpful if a comprehensive database of processing factors and yield factors could be accessed on-line.
- One WG member noted that at present, processing and yield factors are applied at different points in their data systems. They have not been able to simplify the application of these factors although they have identified that moisture loss factors are common between nutrient and processed food recipes and should be able to be unified into a single factor that can be applied across food groupings. Likewise nutrient retention factors and processing factors are the same in concept but are not going to be the same number for any given food/hazard combination. Therefore any system needs to have the capacity to apply various factors on a case-by-case basis and possibly draw on more than one coding system (raw commodities/processed foods) in parallel. Providing support for this through software would, most likely, be more effective than trying to unify the coding systems, if that is the intention.

Additional comments regarding the Codex classification systems

- The Codex classification system and the cluster diets do not have all commodities of interest, particularly for processed food. For example, one WG member noted that they cannot conduct a full assessment of exposure to fumonisins due to the lack of food consumption data for many corn-based products in the cluster diet database. In other cases, some national foods are not included in the Codex food classification systems since both consist mainly of ‘western’ foods
- The Codex classification system is very production oriented and thus it is not feasible to easily link occurrence data to food consumption databases. Additionally, the Codex classification system for processed commodities is mostly limited to food additives issues.
- One WG member suggested that there should be only one code for each food item. In order to distinguish between raw commodities and processed food, the food classification could be structured using facets: a facet to describe the food items, a facet to describe product treatment, another for method of production... For example, the term ‘canned’ may belong to a specific facet, ‘fat free’ to another, etc. This would allow the data provider to enter the additional descriptive information and the data user could use that information to select appropriate foods to use for specific purposes.
- Data providers may already have their own food classification which can be very different from the Codex food classifications used in GEMS/Food. In that case, it, the data provider would have to match its food classification to the GEMS/Food codes. There was general agreement among the WG members that this is a very time-consuming task and that a mapping tool would be helpful.
- It is very time consuming to re-classify all data when you have a lot and the GEMS/Food system is not feasible to capture actual food consumption information.

## 7. DATA ENTRY CONSIDERATIONS

The WG was asked to respond to two questions regarding the data entry/submission process:

- What challenges have been experienced when entering data in the current OPAL I system?
- What would make the data entry process more efficient, for example (but not limited to): drop down menus, ability to copy repeated information, ability to convert national data structure to the GEMS/Food database structure.

Comments and suggestions of the WG follow.

What challenges have been experienced when entering data in the current OPAL I system?

- The difficulty of installing the OPAL software on local computers seems to be a common problem. This is due to either the prohibition of installing 'non-approved' software or differences in operating systems.
- Another frequent comment was that the time required to format and enter data manually into the system was excessive; this is not only tedious but also provides opportunity for data entry errors.
- One country has routinely submitted data in an Excel file rather than via the OPAL software; they had not received feedback from WHO that the Excel format submission required any conversion to GEMS/Food database structure. The data had been easily extracted from the in-house Laboratory Information Management System.
- One member noted that they were never able to get the data import feature in OPAL to work.
- There have been problems when attempting to enter data on foods or contaminants that weren't present on the pick lists. This required contacting WHO to request that missing items be added to the OPAL software, and then waiting for months for an updated revision of OPAL before data entry could continue. Additionally, unless every update is distributed to each contributor, this updating mechanism will result in a number of older versions of the software being used (lack of version control).

What would make the data entry process more efficient?

General comments:

- An important consideration is that incomplete data sets and resource restrictions are a fact of life for many countries. Therefore the GEMS system has to be as simple as possible to use with as few as possible mandatory fields. It should not require countries to obtain additional IT resources simply to upload data into the system.
- It would be useful to provide an Excel file that includes worksheets with the predefined cell format for data submissions and supporting information such as master lists of codes and explanatory notes for data.
- It may be useful to establish a number of 'regional hubs' through which data would be submitted. The regional hub would be responsible for checking the data integrity before forwarding it to the GEMS database.

Drop down menus and auto-fill pick lists:

- Drop down menus are useful for manual data entry but not for upload of large amounts of data at a time (for example from a spreadsheet extracted from our system).
- Auto fill pick lists could be employed to replace the drop down menus.

Use of controlled terminology:

- Controlled text is useful for ensuring accuracy in data entry. However the key thing to consider when specifying text/free form text/controlled language is the ability to readily retrieve information from the database. If exact text matches are not required for the data in a field to be usable then it should not be necessary to control the language that is used.
- Search strategies such as trigram searches and auto-complete functions are easier to use when entering data manually and help to ensure accuracy of data entry.
- One note of caution is that while controlled text and field limits are useful when uploading data manually (i.e. entering a single record at a time), they can be a problem when uploading a large amount of data in a single automated process. When entering data manually it is easy to overlook the restrictions that are in place, which can cause a freeze in the upload process because your data format doesn't match the controlled format, or data to be truncated or lost when extra characters are cut off.
- If controlled terminologies are used, they should be provided in a standard file format. This will allow data provider to develop their own export tools converting their data into the WHO format.

Repeated information:

- The ability to copy (or multiply) entries is a necessity. This is extremely helpful when entering data on multi congener contaminants or analysis with multiple results (e.g. analytical method used for multiple metals).
- For manual uploading of data, a feature could be added that would allow the data provider to indicate the fields that will be repeated from one record to another.
- This information could be submitted as a partially filled upload file.

Automated data transformation:

- Allowing a (semi-) automated transformation of data from one system (e.g. a national database or LIMS) to the GEMS/Food format would be extremely useful. Ultimately, the aim should be to move away from any sort of manual data entry where the data are already digitized in some format.
- One of the most time-consuming aspects of data entry is mapping between the national food classification and the WHO food classification. There are two possibilities for automating this process: 1) Develop a commodities mapping application within the GEMS/Food system, which could be very costly and complex, or 2) convince each country to use the Codex commodities numbering system and map their numbering system to it. The second approach is initially labor intensive but once complete, would help reduce the work load of those responsible for reporting/entering the data. An updated and complete list of Codex commodity codes should be made readily available through the GEMS/Food system.

- The size of the file transmission can be an issue. It may be necessary to create rules for splitting large files into more manageable files (e.g. by month or food) and/or to recommend that transmitted file be zipped.

## 8. RELATED ACTIVITIES AND EXPERIENCES IN OTHER COUNTRIES

Several WG members noted that similar initiatives were being undertaken in their country/region, and related their experiences and approaches. Some of their comments and suggestions were included in previous sections, but some additional information is provided below.

Canada

They are currently in the early process of developing a web-based data upload module designed to circumvent any problems associated with a desktop software data loading. This is in the preliminary planning stage and information gathered could be shared with the WG.

Food Standards Australia New Zealand (FSANZ)

Allowing a (semi-) automated transformation of data from one system (e.g. a national database or LIMS) to the GEMS/Food format would be extremely useful. This type of extract-transform-load process is something that FSANZ is developing in its new data system. (See the comments regarding XML above.) Ultimately, the aim should be to move away from any sort of manual data entry where the data are already digitized in some format.

European Food Safety Authority (EFSA)

EFSA provided suggestions and information throughout this report mainly to give examples of what can be done, and they are willing to participate in any future Working Groups on this topic. Although WHO works on a fully international basis and would not be able to adopt a European system, the documents mentioned below might serve as a starting point for future improvements:

- Much effort has been spent on trying to streamline data capture and reporting. A 'standard sample description model' has been published recently: (http://www.efsa.europa.eu/en/datexcallsfordata/datexsubmitdata.htm). This includes an Excel template with full descriptions of options available and a form with drop down menus (same web reference).
- A 'Data Collection Framework' has been set up for data submissions directly to an Oracle database. It is possible to submit Excel files to the database but the preferred format is xml. To train Member States in using the xml format a pilot project was initiated involving five countries during last year and a new call for five more countries has just been published: (http://www.efsa.europa.eu/en/datex201001/docs/cfpefsadatex201001cp.pdf).
- EFSA is working on data cleaning procedures and the aim is to include the checking during the data submission phase directly during the business-to-business transfer.
- EFSA is also working on the exposure calculation side.

- EFSA recently published a document on how to handle left censored data: (http://www.efsa.europa.eu/en/scdocs/scdoc/1557.htm). This is a perennial problem with many data submissions containing a majority of non-quantified results.
- With regard to food consumption data, rather than using inappropriate production data and sometimes dubious clustering EFSA has moved to actual food consumption data collected at country level. The first implementation of what EFSA called the 'Concise Database' using foods aggregated at very broad levels left a lot to be desired. EFSA has now moved on to the "Comprehensive Database" with foods recorded at a much less aggregated level. They are attempting to standardise also the methodology for food consumption surveys (http://www.efsa.europa.eu/en/press/news/datex091218.htm) and initiate a European attempt to understand what the population in Europe eat (the EUMenu project: http://www.efsa.europa.eu/en/press/news/datex100212.htm).
- Using detailed food consumption data requires the classification of foods using compatible systems both on the consumption and the contaminant/additive occurrence side. Based on European legal requirements including details specified in the pesticide legislation, EuroFIR classification using LanguaL as a template and other sources, EFSA has constructed a preliminary classification system called FOODEX1: (http://www.efsa.europa.eu/en/datex201001/docs/cfpefsadatex201001ax1.xls). It deviates in particular from EuroFIR in the handling of composite foods. As a next step EFSA has formed a Working Group to take this further and if possible harmonise the system so it can be implemented across most EU countries at least as a translational layer (possibly also including translation to GEMS/Food) (http://www.efsa.europa.eu/en/datexwgs/documents/datexfoodclassification.pdf).

United States of America

The US FDA manages a central database into which the FDA labs enter their results. There is no public web-access to this database ; data extractions are done internally and on ad'hoc request. Data for contaminants and pesticides in foods are compiled by FDA's Center for Food Safety and Applied Nutrition (CFSAN) and are made publicly available on the CFSAN website.

The USDA provides web access to nutrient composition data and to the national food consumption data collected in the National Health and Nutrition Examination Survey (NHANES).

## 9. APPENDIX A: GEMS/FOOD DATABASE FIELDS

FOR AGGREGATED DATA

| Field no. | Short field name | Descriptive field name | Field size | Explanatory notes |
|---|---|---|---|---|
| 1 | SN | Serial no. of the record | N8 | Code of centre(N2) + year (N2) + record number (N4; sequential) |
| 2 | CD | Creation date of record | AN11 | Date in format DD(N2)-MMM(A3)-YYYY(N4), e.g. 2-DEC-1998 (leading zero not included) |
| 3 | CC | Country code | A3 | e.g. New Zealand = NEZ |
| 4 | FD | Food identifier | AN6 | Two alphabetic characters (A2) followed by one to four numeric characters (N1-4), e.g. CF2398; see Appendix 3 (following zeros not included) |
| 5 | OR | Food origin | A3 | e.g. New Zealand = NEZ; or blank |
| 6 | SP | Sampling period | N15 | Sampling beginning and end dates in format – month begin/year begin – month end/year end, e.g. 04/1994-06/1995 (leading zeros included) |
| 7 | REP | Sample representativeness | A2 | SW = statistical, whole country; SP = statistical, part country; NW = not statistical, whole country, or; NP = not statistical, part country |
| 8 | NOL | Number of contributing laboratories | N2 | Number of labs for which data have been aggregated (leading zero not included) |
| 9 | AQA | Analytical quality assurance | A2 | IQ = internal quality assurance only; SP = successful proficiency testing, or; OA = officially accredited |
| 10 | CON | Contaminant | N3 | Three digit contaminant code, see Appendix 3 (leading zeros included) |
| 11 | DIM | Dimension of results | A1 | 1 = mg/kg; 2 = μg/kg; 3 = ng/kg, or; R = Bq/kg |
| 12a | LODMIN | LOD minimum | N3.5 | Enter into both LODMIN and LODMAX if one lab |
| 12b | LODMAX | LOD maximum | N3.5 | Enter into both LODMIN and LODMAX if one lab |
| 12c | LOQMIN | LOQ minimum | N3.5 | Enter into both LOQMIN and LOQMAX if one lab |
| 12d | LOQMAX | LOQ maximum | N3.5 | Enter into both LOQMIN and LOQMAX if one lab |
| 13 | BASE | Results based on | A1 | F = fat content; D = dry weight. or; A = as is |
| 14 | N | Number of samples | N5 | Includes all quantified and non-quantified results |
| 15 | < | Number of samples less than the LOQ | N5 | |
| 16a | MIN | Range - minimum | N3.5 | Minimum quantified result |
| 16b | MAX | Range - maximum | N3.5 | Maximum quantified result |
| 17a | X | Mean or best estimate | N3.5 | See Appendix 5 for best estimate if some data are not quantified |
| 17b | $X_L$ | Mean - lower bound | N3.5 | See Appendix 5 if more than 60% of data are not quantified |
| 17c | $X_U$ | Mean - upper bound | N3.5 | See Appendix 5 if more than 60% of data are not quantified |
| 18 | MED | Median or best estimate | N3.5 | See Appendix 5 if more than 50% of data are not quantified |
| 19 | 90th | 90th Percentile | N3.5 | Do not enter value if more than 90% of data are not quantified |
| 20 | STDDEV | Standard deviation | N3.5 | Do not enter value if distribution is not normal or more than 50% of data are not quantified |
| 21 | STATUS | Status of data | N1 | 0 = Not confidential or 1 = Confidential |
| 22 | REM | Remarks/References | AN255 | Information necessary to interpret the data |

FOR INDIVIDUAL DATA

| Field no. | Short field name | Descriptive field name | Field size | Explanatory notes |
|---|---|---|---|---|
| S1 | SN | Serial no. of the sampling unit | N8 | Year (N2) + record number (N6; sequential). Serial number must be unique. |
| S2 | CD | Creation date of record | AN11 | Date in format DD(N2)-MMM(A3)-YYYY(N4), e.g. 2-DEC-1998 (leading zero not included) |
| S3 | CC | Country code | A3 | e.g. New Zealand = NEZ |
| S4 | FD | Food identifier | AN6 | Two alphabetic characters (A2) followed by one to four numeric characters (N1-4). See Appendix 3 of above-mentioned manual or Codex Classification of Foods and Feeds, Codex Alimentarius, Volume 2 Pesticide Residues in Food, 2nd edition, FAO, Rome (1993). For processed food, insert code for the major ingredient and include details on composition and processing in Field 22 Remarks/References. |
| S5 | OR | Food origin | A3 | Contact GEMS/Food Manager for listing; enter country of production, e.g. New Zealand = NEZ; or if both imported and domestic or unknown, leave blank |
| S6 | SD | Sampling date | AN10 | Sampling date in format - day/month/year, e.g. 01/04/1994 (leading zeros included) |
| S7 | REP | Sample reliability | A2 | SW = statistical, whole country; SP = statistical, part country; NW = not statistical, whole country, or; NP = not statistical, part country |
| S8 | LAB | Laboratory identification | N2 | Arbitrary identification number of lab from which data have been collect |
| S9 | AQA | Analytical quality assurance | A2 | IQ = internal quality assurance only; SP = successful proficiency testing, or; OA = officially accredited |
| S10 | STATUS | Status of data | N1 | 0 = Not confidential or 1 = Confidential |
| S11 | REM | Remarks/ References | AN255 | Information necessary to interpret the data |
| C1 | CON | Contaminant | N3 | Three digit contaminant code, see Appendix 3 (leading zeros included). |
| C2 | DIM | Dimension of results | A1 | 1 = mg/kg; 2 = ug/kg; 3 = ng/kg; 4 = pg/kg; or R = Bq/kg. All results on body weight basis. |
| C3 | LOD | Limit of Detection | N3.5 | Limit of Detection for present analysis |
| C4 | LOQ | Limit of Quantification | N3.5 | Limit of Quantification for present analysis |
| C5 | BASE | Results based on | A1 | F = fat content; D = dry weight; C = as consumed or; A = as is. |
| C6 | RESULT | Result | N3.5 | Leave blank if result is not quantifiable |

## 10. APPENDIX B: MEMBERS OF THE WORKING GROUP

**The following ad'hoc experts were involved on the basis of their experience in the WHO Collaborating Centres involved in data submission to the GEMS/Food Programme. Additional experts were invited for their experience as risk assessors and selected from the various regions around the world.**

- Joanne Chan, Singapore ( WHO Collaborating centre SIN17)
- Judy Cunningham, Australia (WHO Collaborating centre AUS38)
- Eloisa Dutra-Caldas, Brazil
- Katie Egan, United States of America (Chair)
- Stefan Fabiansson, European Union (EFSA)
- John Moisey, Canada (WHO Collaborating Centre CAN 64)
- Jiri Ruprich, Czech Republic
- Lydie Soler, France (WHO Collaborating Centre FRA111)
- Yongning Wu, China (WHO collaborating Centre CHN24)

Marie Madeleine Gimou (Cameroon) and Richard Vanhoort (New-Zealand) who were initially invited to participate had declined.

9 789241 500685